CancerNet^{TM} Redistribution via WWW

Gustav Quade, Norbert Püschel, Frederick Far
Institut für Medizinische Statistik, Dokumentation und Datenverarbeitung der
Universität Bonn, Sigmund-Freud-Str. 25, 53105 Bonn

CancerNet from the National Cancer Institute contains nearly 500 ASCII-files, updated monthly, with up-to-date information about cancer and the "Golden Standard" in tumor therapy.

Perl scripts are used to convert these files to HTML-documents. A complex algorithm, using regular expression matching and extensive exception handling, detects headlines, listings and other constructs of the original ASCII-text and converts them into their HTML-counterparts. A table of contents is also created during the process. The resulting files are indexed for full-text search via WAIS.

Building the complete CancerNet WWW redistribution takes less than two hours with a minimum of manual work. For 26,000 requests of information from our service per month the average costs for the worldwide delivery of one document is about 19 cents.

INTRODUCTION

Physician Data Query[1] (PDQ)© was introduced in 1984 by the National Cancer Institute (NCI). The Cancer Information Summaries are parts of PDQ and include:

- Concise summaries for patients and physicians on the prognosis, staging, and treatment of more than 80 major tumor types. Patient information is written in lay language that assists patients in understanding their disease and treatment options.
- Summaries on the assessment and management of commonly encountered problems or conditions associated with cancer and its treatment.
- Detailed summaries of the evidence on screening and prevention for selected types of cancer, including the levels of evidence, significance, and evidence of benefit.
- Summaries on selected investigational or newly approved drugs for the treatment of cancer.
- References and abstracts to key papers in the medical literature.

These cancer information summaries are available through CancerNet. The expertise of more than 100 specialists helps different editorial boards to discuss and modify cancer information statements based on review of recently published data. CancerNet is updated monthly.

CancerNet is available in two languages, English and Spanish. The English version includes nearly 500 ASCII files, each is identified by the string "cn-" and a 6 digit number. Different groups of files are distinguished by the first digit, 1 indicates physician information, 2 indicates patient information etc..

For each CancerNet file the "Contents List" includes the title of the document and the filename, e.g. "cn-400000" for the document "Changes to CancerNet". Before retrieving certain documents for several types of distribution, it is required to access the contents list to find the appropriate document filename.

HISTORY

First Steps

After first experiments with World Wide Web (WWW) in the beginning of 1994 we decided to launch a 100% WWW redistribution of CancerNet to test the suitability of WWW for a complex information database. After signing a Memorandum of Understanding with the NCI, we started to convert the CancerNet files into Hypertext Markup Language (HTML).

To convert the documents into a preformatted HTML-file, we wrote a perl script[2,3] that a) replaced HTML command characters with their escape-sequences and b) put a header and a <pre> in front of and a </pre> at the end of the file. This constituted an exact 1:1 translation of the original ASCII-file. The contents list file from the NCI was divided into several Table of Content (toc) files which were converted

0195—4210/96/$5.00 © 1996 AMIA, Inc.

to HTML manually. An index page with links to the different toc's was created by hand.

After a week all 500 files were converted, the toc's with links to the HTML-documents were built, and our first CancerNet edition with user friendly, topic oriented access to the information became available via WWW. Each document in the WWW space is defined by its Uniform Resource Locator (URL), build from the document type eg. HTTP, the server's internet address, and the name of the file. The URL for the main page of our CancerNet distribution is :

```
http://www.meb.uni-bonn.de/cancernet/
```

In the next version of our scripts we introduced the functionality of building the toc's for the different groups. The effect was to reduce the time for compilation of the monthly CancerNet updates to several hours.

Table of Contents

In the next step we added a table of contents to the beginning of each document. Especially for documents with 30 to 50 pages this was a great improvement, because it allowed the user to navigate just by mouse-click to the right position inside the document.

Text Analysis

Our next objective was to get away from creating preformatted HTML-files and make use of as many HTML-features (like headings, lists, enumerations etc.) as possible. In order to achieve this we completely rewrote our scripts. This "Next Generation" of scripts is still in use and constantly being improved.

CONCEPTS

ASCII to HTML Conversion

The following shows a reduced list of tasks our translation program accomplishes during the translation process:

- Translate HTML command characters and national characters to their HTML escape sequences.
- Detect references to other CancerNet documents and other URLs and convert them to hypertext links.
- Detect headings and enter them into the table of contents.
- Detect text structures like lists, references, tables etc. and convert them to a suitable HTML-representation.
- Detect which part of the original text is running text.
- Preserve as much as possible of the original formatting, especially the indentation of the text.

The translation program works in two phases:

1. The actual translation and generation of the table of contents in a separate file.
2. Merging the translated file and the toc; inserting the value of variables that were not known until after the first phase.

The second phase is rather trivial; in the following we will only discuss the actual translation process that occurs during the first phase.

The basic idea behind the translation is converting sequentially the source text to HTML line by line. (Naturally, a single line of source text may result in the generation of several lines of HTML.) As complex text structures cannot be detected merely by looking at a single line, we implemented a look-ahead of one line and also kept information about previous lines of the source text, most notably a stack of text indentation and type.

Text Indentation with HTML: CancerNet files make heavy use of text indentation as a visual means of structuring the text. We "abuse" HTML glossary lists to produce text indentation in our documents. Glossary lists are a part of HTML-1 and therefore have the advantage of being understood correctly by all webbrowsers.

Character encoding and Hyperlinks: The easiest part of the translation process is replacing the HTML command characters (&, ", < and >) and high ASCII characters with their respective escape sequences, as this can be done via a simple search-and-replace operation. This takes only a few lines of code in perl because of the built-in support for regular expressions in this programming language.

Placing hyperlinks in the text is only marginally more difficult since the search patterns are strings now, for example: "http://", "cn-dddddd", "cs-dddddd" (where d is a single decimal digit).

Here is an excerpt from our translation program that places hyperlinks to CancerNet articles:

```
s/cn-(\d6)/
        <A HREF=\1.html>Art. \1< \/A>/g;
```

Heading Detection: The text is scanned for headings using several criteria:

- Text enclosed within double stars or double dashes
- Text "underlined" with dashes in the following line
- Text all in uppercase
- etc.

A heading is marked with the appropriate HTML tags and, if it is a top-level heading, entered into the table of contents.

A toc-entry consists of a name-tag at the actual occurence of the heading and a hyperlink in the table of contents at the beginning of the document.

Example: The heading **PROGNOSIS** is converted to the following statement:

```
<H2><A NAME="1_PROGNOSIS">
    PROGNOSIS </A></H2>
```

and the corresponding line for the documents toc:

```
<DT><A HREF="#1_PROGNOSIS">
        Prognosis < /A >
```

is written to a temporary file. In the second phase the temporary toc file is inserted at the top of the document.

Running Text: Web browsers format the document text normally in a way that the lines fit into the browsers window, independent of the font size used. However, preformatted text does not always fit into the browser window. In this case the font size has to be changed or the text has to be scrolled for each line from left to right. The difficulty with running text is to keep it apart from other types of text that may not be reformatted by the browser. We use a "guessing" algorithm that gives correct results most of the time and errs only on the safe side. Therefore, in case of doubt, it leaves the orginal format of the text intact.

References: References can be detected by the keyword "References:" and are converted to HTML using numbered lists.

For example:

```
<H3><B>References:</B></H3>
<OL>
<LI>Tarbell NJ, Thompson L, ...
<LI>Marcus KC, Svensson G, ...
</OL>
```

which, using a WWW-browser, looks like:

References:

1. Tarbell NJ, Thompson L, ...
2. Marcus KC, Svensson G, ...

Other Lists: Apart from reference listings, lists in all forms and varieties abound in the CancerNet: Numbered lists with arabian and latin numbers, lists marked with stars, dashes or even "o"s, completely unmarked lists and lists within lists. All these types of lists are recognized by our translation program and converted to HTML.

Note: Numbered lists are not converted to HTML numbered lists. This is due to the fact that the lists in the original CancerNet files are often interrupted by text inserts. As basic HTML numbered lists always start with "1", they cannot be used for these cases.

Other text structures: Since some text structures cannot automatically be detected or meaningfully converted to HTML syntax by our program (eg.: tables, ASCII pictures and some other irregular structures), the original text structures remain. These passages are bracketed by "<PRE>" and "</PRE>" in the source text files. The translation program then copies these text passages verbatim into the HTML doucument.

FULL-TEXT SEARCH VIA WAIS

In addition to our topic oriented access to the CancerNet files we offer a full text search interface using Wide Area Information Server technology (WAIS, z3950) to find certain cancernet files by keywords. We use freeWAIS-0.3 to generate a waisindex for the HTML-documents, excluding the toc files. This index is generated separately for each of the two languages English and Spanish. From the main index of the English and Spanish distribution we provide access to the appropriate HTML based search form. This form allows to enter nested querys using the following boolean search operators: "or", "and", "not". A valid query is e.g.:

```
(cancer and pdq) not breast
```

The query is read by a perl-script running on our server, acting as a local gateway to WAIS. The resulting file names of documents which correspond to the search statement are converted to standard HTML links and displayed by the user's browser. This interface is also used to give clients access to our CancerLit redistribution, which includes the cancer-related literature of the last six months.

SURVEY

From December 12th, 1995 until January 21th, 1996 we carried out a user survey. From every document we offered a link to a HTTP form, which could be filled out by the user. All answers are collected and analysed using the Statistical Analysing System (SAS)[4].

RESULTS

For each document delivered to a browser, our server writes information about the requesting hosts address, the browser used, date and time of delivery, the name of the requested file, and the number of bytes delivered into the access log file. The URL of the document which offers the link to one of our CancerNet documents is logged in a reference log file.

From analysis of our servers log files with SAS we know that our service, which was the world's first 100% HTML version of CancerNet, is frequently used by users around the world. Starting with 131 accesses and about 1 Mbyte of data sent in May 1994 we reached 40522 requests with an amount of data sent all over the world of 953 Mbytes in February 1996. Since the fall of 1994 our service has been growing more than 20% per month. The following table shows usage statistics for selected months. All values are including the accesses to the tocs but without image loading.

Date	Mbytes	# Accesses
May 94	1	131
Dec 94	11	1059
Mar 95	58	4436
Jun 95	119	6798
Sep 95	331	17092
Dec 95	451	25524
Feb 96	953	40522

With the exception of 15 minutes on January 8th, 1996 (harddisk failure and rebooting 3 times), the service was available. Due to bottlenecks in US networks for international data traffic users from the US had problems reaching our server for serveral time periods in the past. These problems have been solved since January 1st, 1996.

In the past two years users from over 60 countrys requested information from our server. Today over 75% of requests originate from the US. Starting in the summer of 1995 we noticed a remarkable increase of requests from users using an online provider like AOL, Compuserve, Prodigy or the Deutsche Telekom to access the internet. This is due to the fact that the online providers offered internet access via a WWW browser to their customers and set up subject oriented lists with links to worthwhile services to start with. The following table shows the main online providers and the percentage of CancerNet usage by their customers in January 1996.

Provider	Mb	# Accesses
AOL	116	4585 11.3%
Compuserve	22	916 2.3%
Prodigy	21	824 2.0%
Deutsche Telekom	5	216 .5%

Analysing the reference log file we found that the majority of initial accesses to our documents were done from links provided by so-called internet search engines. These search engines retrieve all accessible documents in the internet, build up a full text index from the document's content and offer access to the document's URLs via a keyword search interface. How many initial accesses to our CancerNet documents were occured from links provided by the different search engines is shown in the following table. Internal links are excluded.

Engine	References
Yahoo	22.4%
Webcrawler	20.2%
Altavista	7.2%
Infoseek	4.8%
Lycos	3.7%
Opentext	2.4%

Several hundred servers around the world, including organisations like the WHO, provide links to our CancerNet service for their customers, many of them in form of guides to health related information services. The health related guides which acted most as mediator to our service are:

Guide to Service	Ref.
AOL	4.8%
Karolinska Institute (Sweden)	2.4%
Arnes (Slowenia)	1.7%
Martindale's Health Guide	1.4%

The form for our survey was filled out by 538 users from 38 different countrys. About 75% of these users are living in North America, the gender rate is 62.3% males to 37.7% females. The average age was 39.4 years for female and 43.0 for male users. Nearly 60% of all requests on our CancerNet service were done in connection with a patient's disease. About 20% of the survey participiants were physicians. The question about the rating of our service was answered by 46.9% with excellent, 46.9% with good, 5.7% with medium, and 0.5% with poor.

Costs: For maintenance of our server und the monthly updating we allow for about 3 to 4 days of work per month. This work is done mainly by students. The average costs for labor per month are about $700. A monthly quota of one Gbyte worldwide data traffic over a 2Mbit line costs about $3500 in Germany. Including the expense for amortisation of our investments in hardware we delivered in January about 26500 documents (tocs excluded) at a total cost of $5000. This amounts to about 19 cents for a worldwide delivery of one document.

DISCUSSION

Our experience with two years of providing access to CancerNet demonstrates that our perl scripts in combination with the Web technology allow for the construction of a worldwide accessable information database. The great variety in the CancerNet layout contributed to increased effort in programming. Even then not all parts of the original CancerNet files could be transformed into HTML style documents without some manual work.

The current HTML standard does not cover all layout styles used in the CancerNet files. The server technology allowed a nearly uninterrupted access to our service but the international connections were not stable enough and their capacity was too low for certain time periods in the past.

CONCLUSION

Using the Web technology it is possible to provide worldwide access to information on high quality guidelines and standards. Patients and physicians can be informed with a minimum of costs. This is a key point in improving the quality of health care in the western countries as well as for economies in transition.

Acknowledgments

We are grateful to the NCI which allowed us to redistribute the CancerNet database.

References:

1. Hubbard SM, Martin NB, Thurn AL, NCI's Cancer Information Systems – Bringing Medical Knowledge to Clinicians. *Oncology* 9(4): 302-313, 1995.

2. Wall L, Schwartz RL, Programming Perl, O'Reilly & Associates, Inc., März 92

3. Schwartz RL, Learning Perl, O'Reilly & Associates, Inc., Nov. 93

4. SAS Institute Inc., SAS Language and Procedures: Usage, Version 6, First Edition Cary, NC: SAS Institute Inc., 1989. 638 pp.